Fine-Tuning Nonhomogeneous Regression for Probabilistic Precipitation Forecasts: Unanimous Predictions, Heavy Tails, and Link Functions

Gebetsberger, Manuel; Messner, Jakob W.; Mayr, Georg J.; Zeileis, Achim

Published in:
Monthly Weather Review

Link to article, DOI:
10.1175/MWR-D-16-0388.1

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Fine-Tuning Nonhomogeneous Regression for Probabilistic Precipitation Forecasts: Unanimous Predictions, Heavy Tails, and Link Functions

MANUEL GEBETSBERGER

Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

JAKOB W. MESSNER

Department of Statistics, University of Innsbruck, Innsbruck, Austria

GEORG J. MAYR

Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria

ACHIM ZEILEIS

Department of Statistics, University of Innsbruck, Innsbruck, Austria

(Manuscript received 10 October 2016, in final form 14 September 2017)

ABSTRACT

Raw ensemble forecasts of precipitation amounts and their forecast uncertainty have large errors, especially in mountainous regions where the modeled topography in the numerical weather prediction model and real topography differ most. Therefore, statistical postprocessing is typically applied to obtain automatically corrected weather forecasts. This study applies the nonhomogenous regression framework as a state-of-the-art ensemble postprocessing technique to predict a full forecast distribution and improves its forecast performance with three statistical refinements. First of all, a novel split-type approach effectively accounts for unanimous zero precipitation predictions of the global ensemble model of the ECMWF. Additionally, the statistical model uses a censored logistic distribution to deal with the heavy tails of precipitation amounts. Finally, it is investigated which are the most suitable link functions for the optimization of regression coefficients for the scale parameter. These three refinements are tested for 10 stations in a small area of the European Alps for lead times from +24 to +144 h and accumulation periods of 24 and 6 h. Together, they improve probabilistic forecasts for precipitation amounts as well as the probability of precipitation events over the default postprocessing method. The improvements are largest for the shorter accumulation periods and shorter lead times, where the information of unanimous ensemble predictions is more important.

1. Introduction

Physically based ensemble forecasts have become the standard in operational weather forecasting to capture atmospheric forecast uncertainty (Leith 1974). Slightly perturbed initial conditions and/or different model formulations are used to derive an ensemble of numerical weather predictions. If the different initial conditions and model formulations reflect the initial condition and model uncertainty this ensemble should reflect the forecast uncertainty. However, because not all error sources can be considered these ensembles are often still biased and underdispersive (Hamill and Colucci 1998; Mullen and Buizza 2002; Bauer et al. 2015).

The European Alps represent a region with an extraordinarily complex topography. Unresolved valleys and mountain ridges cause missing local effects and distort precipitation patterns and amounts. Most of the

Denotes content that is immediately available upon publication as open access.

Corresponding author: Manuel Gebetsberger, manuel.gebetsberger@uibk.ac.at

DOI: 10.1175/MWR-D-16-0388.1

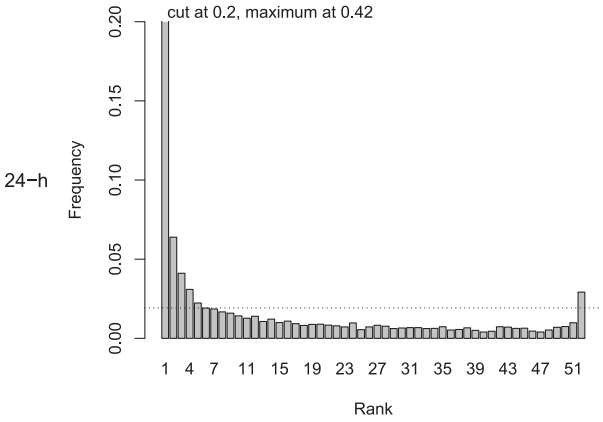

FIG. 1. Rank histogram for 24-h precipitation sums from +24 to +48 h based on raw data from the 51-member ECMWF ensemble forecasts, evaluated for 10 stations located in North Tyrol (Austria) and South Tyrol (Italy): the *x* axis denotes the rank (1–52) and the *y* axis denotes the observed frequency at this particular rank. Dotted horizontal line indicates perfect calibration at 0.02.

precipitation is rained out before it reaches inner alpine valleys, leading to drying ratios of about 35% (Smith et al. 2003).

Therefore, systematic errors and underdispersion are very pronounced for this region, as illustrated by the rank histogram (Hamill and Colucci 1998; Anderson 1996; Talagrand et al. 1997) in Fig. 1. Data used to create this figure are based on three years of observed precipitation amounts and ECMWF ensemble forecasts for multiple stations in this region (see section 3a for details). The rank histogram in Fig. 1 highlights a strong bias with a peak at rank 1, where precipitation amounts are strongly overestimated by the raw ensemble. Additionally, an underdisperison is visible since observations are mostly below the lowest and above the highest member forecast (on rank 1 and 52).

To correct for these errors and to supply automatically corrected forecasts to weather services, the raw ensemble is often statistically postprocessed. For probabilistic quantitative precipitation forecasts various nonparametric (Krzysztofowicz and Sigrest 1999; Hamill et al. 2015; Zhu and Luo 2015) and parametric (Roulston and Smith 2003; Gneiting et al. 2005; Raftery et al. 2005; Sloughter et al. 2007; Wilks 2009) approaches have been proposed. Wilks (2011) and Wilks and Hamill (2007) compared some of these but could not identify one single best method.

Hence, this study focuses on the widely used strategy that is known as ensemble model output statistics (EMOS) or nonhomogeneous regression (NHR) approach (Gneiting et al. 2005). This approach has been tested extensively for different variables (e.g., temperature, mean sea level pressure, wind, and precipitation), and appropriate distributions: Gaussian (Gneiting et al. 2005; Feldmann et al. 2015), truncated normal (Thorarinsdottir and Gneiting 2010), gamma (Wilks 1990), generalized extreme value (GEV; Scheuerer 2014), or censored Gaussian and logistic (Wilks 2009; Messner et al. 2014a, b; Stauffer et al. 2017a). Gneiting and Katzfuss (2014) review suitable distributions for certain variables, statistical ensemble postprocessing, and verification techniques in general.

Apart from different distributions, various other extensions have been proposed to improve the classical NHR such as including neighborhood information to address displacements errors (Theis et al. 2005; Ben Bouallègue and Theis 2014; Scheuerer 2014), accounting for spatial forecast correlations (Feldmann et al. 2015), additional covariates covering seasonal or annual variations (e.g., Stauffer et al. 2017a), or differently weighted NWP models (e.g., Hemri et al. 2016) to account for NWP performance differences.

However, almost all of these extensions need to acquire additional input data [e.g., additional grid points (Scheuerer 2014) or different NWP models (Hemri et al. 2016)]. In this study we present and discuss three purely statistical refinements to improve NHR precipitation forecasts that do not require any additional input data. Clearly, these refinements can also be combined with other extensions such as the ones listed above.

Usually, NHR uses only the (weighted) ensemble mean and standard deviation as regressor variables. However, Sloughter et al. (2007), Bentzien and Friederichs (2012), and Scheuerer (2014) found improvements for precipitation forecasts by additionally using the fraction of zero ensemble members. Starting out from this idea, we argue that for our study area it is not so natural to capture the influence of this zero fraction by a *linear* regressor because in our data unanimous zero precipitation ensemble predictions (i.e., where none of the members predicts any precipitation) almost always correspond to dry anticyclonic situations without any observed precipitation. To exploit this finding, we propose a split approach for NHR that switches to a different parameter set for these unanimous ensemble forecasts and provides much sharper forecast distributions.

Furthermore, a Gaussian parametric distribution is not appropriate for precipitation data that have a physical limit at zero. This limit can be incorporated by censoring the distribution (Cohen 1959), but the tail of events with large precipitation amounts is often also underestimated by the Gaussian distribution. To overcome this, our second statistical refinement uses a heavy-tailed distribution that deals with these precipitation characteristics, similar to Scheuerer (2014) or Scheuerer and Hamill (2015).

Additionally, the nonnegativity of dispersion parameter of the distribution has to be ensured in the cause of the numerical optimization of regression coefficients. The literature describes two solutions to fulfill this requirement: squaring the optimization value (Gneiting et al. 2005) or applying a link function to the dispersion submodel (Messner et al. 2014a). A comparison of these concepts has not been made so far and will be performed in this study as the third statistical refinement.

This article is structured as follows. The statistically motivated refinements are introduced in detail in section 2. Section 3 describes the study area and comparison setup. Section 4 presents our results, which will be summarized in section 5 with some concluding remarks.

2. Refinements

In this section we briefly describe the basic NHR framework, followed by our three statistical refinements: split approach for unanimous predictions, heavy tails, and link functions.

a. Nonhomogeneous regression

NHR defines one type of distributional regression models (Klein et al. 2015) and was initially developed for a Gaussian response such as temperature (Gneiting et al. 2005). Two linear equations correct for the location part [Eq. (1)] and the scale part [Eq. (2)], respectively. Typically, the Gaussian parameters for location and scale (μ_i, σ_i) are linearly linked to the NWP ensemble mean ($\overline{\text{ens}_i}$) and ensemble standard deviation ($\text{SD}_{\text{ens},i}$) for each event i:

$$\mu_i = \beta_0 + \beta_1 \times \overline{\text{ens}_i}, \tag{1}$$

$$\sigma_i = \gamma_0 + \gamma_1 \times \text{SD}_{\text{ens},i}. \tag{2}$$

The four coefficients (β_0, β_1, γ_0, γ_1) can be estimated simultaneously by numerically optimizing the log-likelihood function:

$$\text{logLik} = \sum_{i=1}^{N} \log[f(\text{precip}_i)], \tag{3}$$

which is defined as the sum over the logarithmic densities of the probability density function (PDF) f, evaluated at the observed value precip_i. In the classical NHR approach, f is the Gaussian PDF.

Since precipitation data are nonnegatively defined and skewed to the right, this Gaussian NHR has to be modified. The simple approach of censoring the distribution at a certain threshold (Cohen 1959) was found to be effective for precipitation amounts (Messner et al. 2014a; Stauffer et al. 2017a). This threshold is typically

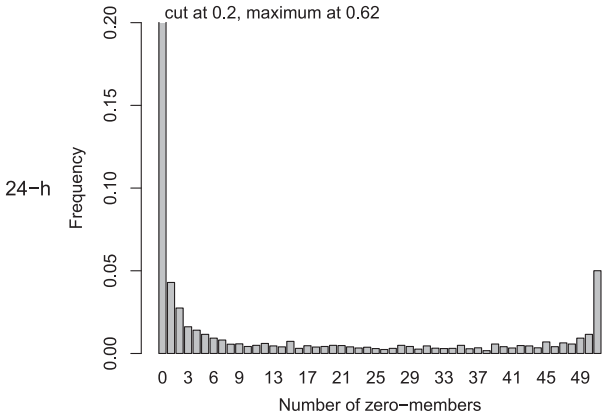

FIG. 2. Frequency of the 51-member ECMWF ensemble forecasts containing a certain number of members being zero (0–51), evaluated for 10 stations located in North Tyrol (Austria) and South Tyrol (Italy) for 24-h sums from +24 to +48 h: the x axis denotes the number of members being zero and the y axis denotes the frequency.

defined at zero for precipitation. One assumes a latent Gaussian process y, which is allowed to become negative but is censored at zero to obtain sensible values for precipitation:

$$\text{precip}_i = \begin{cases} 0 & y_i \le 0 \\ y_i & y_i > 0 \end{cases}. \tag{4}$$

If the latent process becomes positive and far away from zero, the effect of censoring vanishes and the censored Gaussian distribution leads to the Gaussian distribution.

The log-likelihood function, which has to be optimized, differs from Eq. (3) by distinguishing between events on the censoring level ($\text{precip}_i = 0$) and above the censoring level ($\text{precip}_i > 0$):

$$\text{logLik}_i = \begin{cases} \log[F(0)] & \text{precip}_i = 0 \\ \log[f(\text{precip}_i)] & \text{precip}_i > 0 \end{cases}, \tag{5}$$

where F represents the cumulative distribution function (CDF) and f the PDF, evaluated at the censoring level zero or the observed value precip_i, respectively.

b. Split approach

In the introduction we have already implied the importance of using the fraction of ensemble members being zero. Scheuerer (2014) already used this information for probabilistic precipitation forecasts in Germany. Adding a new regressor variable frac into the location part of Eq. (1), which accounts for the fraction of K members without precipitation improved the forecasts.

This fraction is illustrated in Fig. 2 for 10 alpine stations in North and South Tyrol (see section 3a for

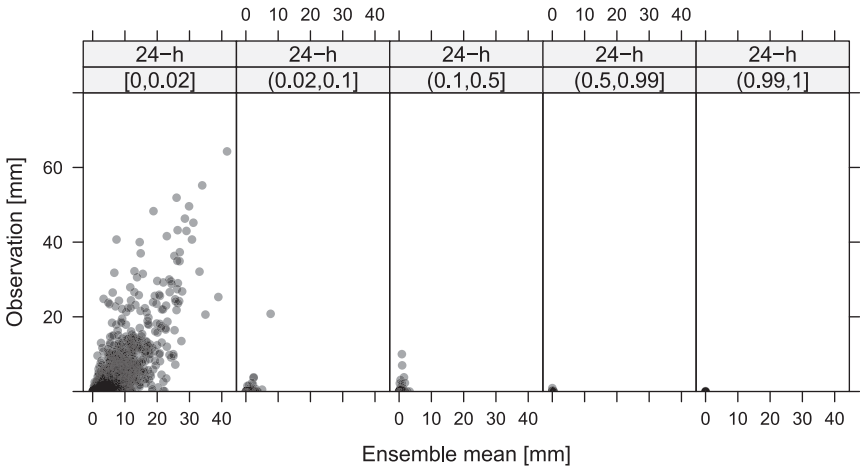

FIG. 3. The 24-h precipitation sums of +48-h ensemble mean ECMWF forecast against observed precipitation at station Innsbruck. Columns further show cases conditional on the fraction of the 51 EPS members without precipitation (0–0.02, 0.02–0.1, 0.1–0.5, 0.5–0.99, and 0.99–1). Darker shading of points indicates more events.

details) for the used 51-member ECMWF ensemble. Most frequently all ensemble members unanimously have precipitation (peak at 0) or all members unanimously have no precipitation (peak at 51). Intermediate numbers of 1–50 members predicting precipitation occur less frequently. Therefore, if this pattern matches (or at least correlates with) the (lack of) observed precipitation in nature, it is possible to improve the forecasting skill in the situations with (almost) unanimous zero predictions from the ensemble.

As an example for a general pattern, Fig. 3 shows the ensemble mean value against the observed precipitation amount to be conditional on the fraction of members forecasting no precipitation. Data are shown for the city of Innsbruck, Austria, and daily precipitation amounts within the available data period, given the 51-member ensemble of the ECMWF. The forecasts of all ensemble members predicting no precipitation (fraction larger than 0.99) are unanimous in the sense that no precipitation has been observed. This figure also illustrates, that such unanimous cases become imperfect when looking on lower levels for the fraction of zeros where precipitation is observed (e.g., fraction larger than 0.1 and smaller than 0.5).

Hence, Fig. 3 suggests that if (nearly) all ensemble members are zero and frac is (close to) 1, no (or only little) precipitation occurs and the regression relationship almost collapses. To improve the performance of the approach of Scheuerer (2014) in our region of interest (section 3a), we propose to use an interaction term instead, which can also be interpreted as splitting the data at a certain split level ν and will be referred to as "split approach."

This split approach uses a binary variable z_i to indicate whether (almost) all ensemble members are zero:

$$z_i = \begin{cases} 1 & \text{if} \quad \text{frac}_i \geq \nu \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

An obvious split level is $\nu = 1$ when *all* ensemble members unanimously forecast no precipitation. However, relaxing the split level to lower values might also be useful (see below).

This new regressor enters the NHR equations as an interaction term:

$$\mu_i = \beta_0 + \beta_1 \times \overline{\text{ens}_i} \times (1 - z_i) + \beta_2 \times z_i, \tag{7}$$

$$\sigma_i = \gamma_0 + \gamma_1 \times \text{SD}_{\text{ens},i} \times (1 - z_i), \tag{8}$$

which can be interpreted as follows: the usual censored NHR with slopes β_1 and γ_1, respectively, is only estimated when a large fraction $(1 - \nu)$ of ensemble members predict precipitation (i.e., $z_i = 0$). Conversely, for $z_i = 1$, when (almost) all ensemble members have no precipitation, the regression collapses to the climatological values for mean $\mu_i = \beta_0 + \beta_2$ and standard deviation $\sigma_i = \gamma_0$.

Typically, the coefficient β_2 will be negative leading to lower predicted precipitation. Because of the censoring, the probability for positive precipitation may become arbitrarily small if β_2 becomes increasingly negative. For this reason $\sigma_i = \gamma_0$ is also kept fixed to avoid that both mean and standard deviation collapse to zero.

The choice for the "best" split point between the NHR regression and simple climatological fit is not as obvious as it may seem. Considering Fig. 2, $\nu = 1$ seems

to be sufficient because there are only few observations with large fractions but below 1. However, from Fig. 3 for station Innsbruck it might also be reasonable to switch to a lower value of $\nu = 0.5$. This value might even be lower at different stations and lead times.

c. Heavy tails

Although the censored Gaussian distribution is able to capture precipitation characteristics (nonnegativity, many observations at zero), distributions exist that better describe rare events with large amounts of precipitation.

We selected the censored logistic distribution, which has a heavier tail than the Gaussian. The censored logistic distribution was found to be useful for both short accumulation periods of 24-h sums (Messner et al. 2014a,b) and longer ones of 6 days (Wilks 2009). Shorter accumulation periods than 24-h sums show similar characteristics of precipitation (nonnegativity, observations at zero) to that of longer accumulation periods, except for a higher frequency of zero precipitation events. Clearly, as accumulation periods become much longer (weekly or monthly) fewer events without precipitation occur so that the effect of censoring decreases.

Censoring and log-likelihood maximization can then be performed as before, but using the logistic PDF [Eq. (9)] and CDF [Eq. (10)] as follows:

$$f(y,\mu,\sigma) = \frac{e^{-(y-\mu)/\sigma}}{\sigma \times (1+e^{-(y-\mu)/\sigma})^2}, \tag{9}$$

$$F(y,\mu,\sigma) = \frac{1}{1+e^{-(y-\mu)/\sigma}}. \tag{10}$$

Note that in order to be consistent with the censored NHR framework of Eqs. (7) and (8), σ defines the scale parameter and μ defines the location parameter of the logistic distribution.

Clearly, there might be other suitable distributions accounting for rare events. Reasonable results for precipitation data have also been achieved with GEV distribution (Scheuerer 2014) over Germany, and the censored shifted Gamma distribution (Scheuerer and Hamill 2015) over the United States. These two distributions can have an even more pronounced tail than the censored logistic distribution. The best choice will depend on the region and accumulation period.

d. Link functions

Since the scale parameter in Eq. (8) is nonnegatively defined, we have to ensure that individual predictions are kept nonnegative during log-likelihood optimization. This can be achieved in two ways: by parameter constraints for γ_0, γ_1 (e.g., squaring these coefficients; Gneiting et al. 2005), or by using a suitable link function (e.g., log link; Messner et al. 2014b). We will investigate differences in forecast skill from using three different link functions g for the scale submodel:

$$g(\sigma) = \gamma_0 + \gamma_1 \times g(\mathrm{SD}_{\mathrm{ens}}) \times (1 - z). \tag{11}$$

They are the following:

- quadratic (quad): $g(\sigma) = \sigma^2$ (Gneiting et al. 2005).
- logarithmic (log) $g(\sigma) = \log(\sigma)$ (Messner et al. 2014b).
- identity (id): $g(\sigma) = \sigma$ (Scheuerer 2014).

These three link functions will be applied in conjunction with our split approach.

3. Data and setup

This section defines the data for our research area and the comparison setup for the statistical models.

a. Data

As mentioned in the introduction, raw ensemble forecasts for precipitation amounts suffer from large bias and dispersion errors (Fig. 1) for our mountainous region of interest. This region is located in the areas of North Tyrol (Austria) and South Tyrol (Italy) and embedded in a complex environment of the central European Alps (Fig. 4). It is famous for wine and fruit growing, where precipitation events and precipitation amounts can strongly influence the evolution of plant pathogens (Löpmeier et al. 2012; Carisse et al. 2009).

The ensemble forecasts are from the operational ensemble prediction system (EPS) of the ECMWF, with a horizontal grid size of 32 km. Its 51 members are taken to be exchangeable and yield a discrete forecast distribution. Direct model output is bilinearly interpolated to 10 station sites of interest using the 4 nearest grid points and precipitation is aggregated over periods of 6 and 24 h, respectively. This low horizontal resolution does not reflect the real topography, so that the spatial variability of the raw ensemble is much lower than the observed variability of precipitation patterns (e.g., Stauffer et al. 2017b).

Observed precipitation amounts are from 10-min measurements of automated weather stations, which are owned by the local weather services. Datasets cover the period from 1 January 2011to 1 January 2014 for the stations in South Tyrol, and 11 January 2011 to 31 January 2017 for Innsbruck in North Tyrol. These datasets are also used to calculate the 97% quantiles for high precipitation amounts. Values range from 12.9 to 22.7 mm for 24-h sums, and from 2.9 to 8.7 mm for 6-h sums.

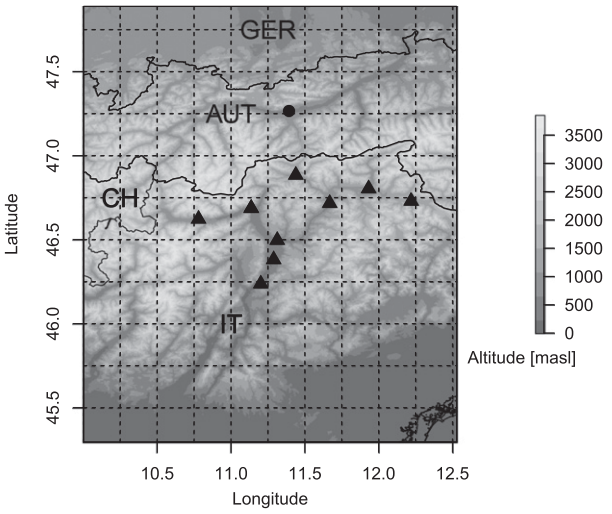

FIG. 4. Used stations within the region of interest [Austria (AUT), Italy (IT), Switzerland (CH), and Germany (GER)]. The filled circle displays the station at Innsbruck. The dotted grid illustrates the underlying horizontal grid size of the 51-member ECMWF ensemble forecasts.

Availability of forecast/observation pairs allows the analysis of the second forecast day (+24 to +48 h), except for Innsbruck where EPS forecasts are available to +144 h. All forecasts are from the 0000 UTC run of the ECMWF EPS.

b. Comparison setup

Table 1 gives an overview about the statistical models used in this article. To quantify the quality of our new split approach, we will use a reference model. The reference approach defines the censored Gaussian NGR and uses the quadratic link with a parameter constraint for the estimated scale coefficients (quad), as used by Gneiting et al. (2005) for temperature forecasts. This model is extended by using the fraction of members being zero (quad_frac), as proposed by Scheuerer (2014). Finally, we use our split approach with the quadratic-link (quad_split), the logarithmic-link (log_split), and the identity-link (id_split). Except for the log-link split model, all models use the parameter constraint of squaring the coefficients in the scale submodel.

The optimization of those models is performed in R with the package *crch*, which performs maximum likelihood optimization (R Core Team 2016; Messner et al. 2016).

To have a fair comparison, performance measures are computed with a tenfold cross validation as in Messner et al. (2014a). Datasets (individual cases) are divided for each station and lead time separately into 10 blocks of approximately the same length by randomly selecting subsamples. Each block is predicted with models trained on the remaining nine-tenths of data. Thus independent forecasts (test data) for the whole period are available to compute verification measures [e.g., continuous ranked probability score (CRPS)] for each event. Averages over these scores are either derived directly on the test data once, or in case for the evaluation of lead time performance, a bootstrap approach is used to estimate the sampling distribution of these averages. Therefore 500 averages are derived for 500 random samples with replacement of the individual scores.

TABLE 1. Overview of statistical models used for comparison: zero information describes whether the fraction of members being zero (frac) or the split approach with the z regressor are used.

Model name	Zero information	Link function in scale submodel	Parameter constraint on γ_0, γ_1
quad	—	Quadratic	Nonnegative
quad_frac	frac	Quadratic	Nonnegative
quad_split	z	Quadratic	Nonnegative
log_split	z	Logarithmic	—
id_split	z	Identity	Nonnegative

As already indicated, model performance is evaluated on the CRPS (Hersbach 2000; Gneiting et al. 2005; Wilks 2011):

$$\mathrm{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}[F_i(x) - H_i(x - y_i)]^2\,dx, \qquad (12)$$

where F_i defines the forecasted CDF and $H_i(x - y_i)$ the Heaviside function, which takes the value 0 if $x < y_i$ and 1 otherwise. This squared difference between the forecasted CDF and the Heaviside function evaluated at the observed value y_i is integrated over the real axis x for each event, and further averaged over the number of n events. The CRPS achieves zero at best, and can diverge to $+\infty$ in the worst case.

To compare the performance of different statistical models (Table 1), we further compute the continuous ranked probability skill score (CRPSS):

$$\mathrm{CRPSS} = 1 - \mathrm{CRPS}_{\mathrm{mod}}/\mathrm{CRPS}_{\mathrm{ref}}, \qquad (13)$$

where $\mathrm{CRPS}_{\mathrm{mod}}$ is each model score and $\mathrm{CRPS}_{\mathrm{ref}}$ is our reference approach. A positive CRPSS indicates better skill than the reference.

Furthermore, forecasts for probability of precipitation (PoP; amounts >0 mm) and occurrence of high precipitation amounts (PoP, amounts > climatological 97% quantile) are analyzed by the Brier score (BS; Brier 1950):

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2, \qquad (14)$$

TABLE 2. Median values of CRPS as a metric for full distribution forecasts, and Brier scores (BS) as metric for the exceedance of two thresholds of precipitation amounts: 0 mm (BS PoP) and the observed 97% quantile at each forecast site (BS high). Analysis for the raw ensemble, Gaussian, and logistic models, evaluated separately for different accumulation periods (24; 6 h) of all stations for forecast day 2 (+ 24 to + 48 h).

		CRPS		BS PoP		BS high	
Type	Name	24 h	6 h	24 h	6 h	24 h	6 h
Raw ensemble	EPS	1.6137	0.5870	0.4246	0.4356	0.0198	0.0216
Gaussian models	quad	1.2105	0.4065	0.1071	0.0868	0.0188	0.0232
	quad_frac	1.2019	0.4006	0.1062	0.0858	0.0187	0.0227
	quad_split	1.2091	0.4035	0.1069	0.0868	0.0187	0.0231
	log_split	1.1922	0.3984	0.1072	0.0853	0.0183	0.0226
	id_split	1.1967	0.3995	0.1059	0.0861	0.0185	0.0230
Logistic models	quad	1.1930	0.4003	0.1022	0.0842	0.0189	0.0225
	quad_frac	1.1923	0.3992	0.1021	0.0837	0.0190	0.0224
	quad_split	1.1928	0.3996	0.1021	0.0842	0.0189	0.0225
	log_split	1.1930	0.3987	0.1025	0.0831	0.0188	0.0220
	id_split	1.1905	0.3988	0.1014	0.0844	0.0189	0.0223

which is a mean squared difference between the forecast probabilities p_i and the binary value of precipitation yes or no o_i. Herein, i defines the index for single events and n is the number of events used for evaluation. Hence, BS is between zero (best) and one (worst). For high precipitation amounts we define the threshold as the site-specific observed 97% quantile for different accumulation periods. This is performed in order to share the same climatological event frequency in the study area instead of choosing a fixed threshold (Hamill and Juras 2006). Additionally, Brier skill scores (BSS) are computed as in Eq. (13) by using BS instead of CRPS.

4. Results

This section is structured as followed: first, we will briefly compare the statistical models to the raw ensemble, followed by the quantification of our three statistical refinements against the reference postprocessing method.

a. Comparison to raw ensemble forecasts

It is essential that postprocessing has to improve the raw ensemble forecasts. We therefore perform a brief ensemble evaluation with the CRPS for the probabilistic forecasts, and the Brier score to check probability forecasts for certain thresholds, both described in the following.

Although the ensemble does not provide a full continuous probability distribution, it is possible to verify the empirical CDF following Hersbach (2000). Additionally, the fraction of ensemble members predicting precipitation can be used to verify the PoP. Average CRPS and BS values are summarized in Table 2, which displays median values taken over the individual cases. Corresponding skill scores computing a measure for improvement against the raw ensemble are based on the values of Table 2 and are provided in Table 3.

Clearly, censored Gaussian and censored logistic models show lower CRPS values than the raw ensemble, both improving the raw forecasts by a value of about 26% and 32% (24- and 6-h sums, respectively). The CRPS is generally smaller for 6-h sums, since smaller precipitation amounts are observed more frequently than for 24-h sums.

Regarding the PoP, the raw ensemble could also clearly be improved by all statistical models. The 24-h sums obtain a Brier score of 0.42 on median, and 6-h sums a score of 0.44 on median. Compared to the raw ensemble, the postprocessed forecasts of all statistical models improve by about 76% and 80%, respectively.

TABLE 3. Skill scores (in %) for the median verification measures shown in Table 2: continuous ranked probability skill score (CRPSS) and Brier skill score (BSS) for the precipitation thresholds of 0 mm (BSS PoP) and the 97% quantile (BSS high), shown for accumulation periods of 24- and 6-h sums. Improvement (skill) is shown against the raw ensemble, which has no skill against itself.

		CRPSS		BSS PoP		BSS high	
Type	Name	24 h	6 h	24 h	6 h	24 h	6 h
Raw ensemble	EPS	0.0	0.0	0.0	0.0	0.0	0.0
Gaussian models	quad	25.0	30.8	75.8	79.5	3.1	3.7
	quad_frac	25.5	31.8	76.0	79.7	3.4	5.8
	quad_split	25.1	31.3	75.8	79.5	3.3	4.0
	log_split	26.1	32.1	75.7	79.9	5.6	5.9
	id_split	25.8	31.9	76.0	79.7	4.7	4.5
Logistic models	quad	26.1	31.8	76.9	80.1	2.2	6.3
	quad_frac	26.1	32.0	76.9	80.2	2.1	7.0
	quad_split	26.1	31.9	76.9	80.1	2.3	6.4
	log_split	26.1	32.1	76.8	80.4	2.9	8.5
	id_split	26.1	32.1	76.8	80.4	2.9	8.5

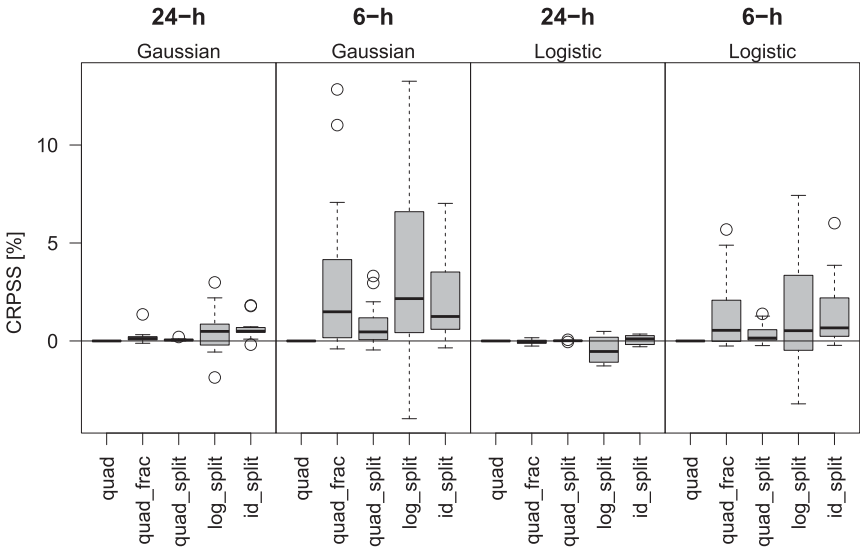

FIG. 5. CRPSS for censored Gaussian and logistic models (Table 1). Reference is the quad model without splitting. Results from tenfold cross validation, where each value represents individual lead times for 10 stations, are shown for different accumulation periods (24 and 6 h) between +24 and +48 h. Boxplots include the interquartile range (0.25–0.75) shown in gray boxes, whiskers show the ±1.5 times interquartile range, and outliers are shown in solid circles.

As the ensemble probability is given by the fraction of members being nonzero, the high BS values indicate that the ensemble strongly overestimates precipitation occurrence. This also adds to Fig. 2 where 24-h PoP is forecasted by 100% nonzero members in 62% of the events (peak at 0). Additionally, the intense peak on rank 1 in Fig. 1 highlights that a large number of observed values are below the lowest member forecast. If a large number of members predict precipitation, ensemble ''probabilities'' for precipitation occurrence become

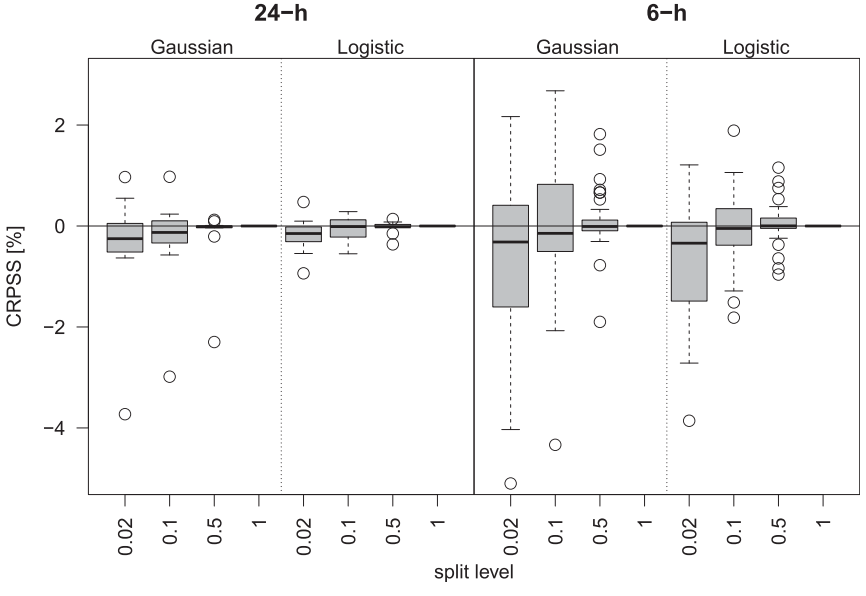

FIG. 6. CRPSS for censored Gaussian and logistic split models using the log link for different split levels. The reference model is split at $\nu = 1$. Results from tenfold cross validation, where each value represents individual lead times for 10 stations, are shown for different accumulation periods (24 and 6 h) between +24 and +48 h. Boxplots are as in Fig. 5.

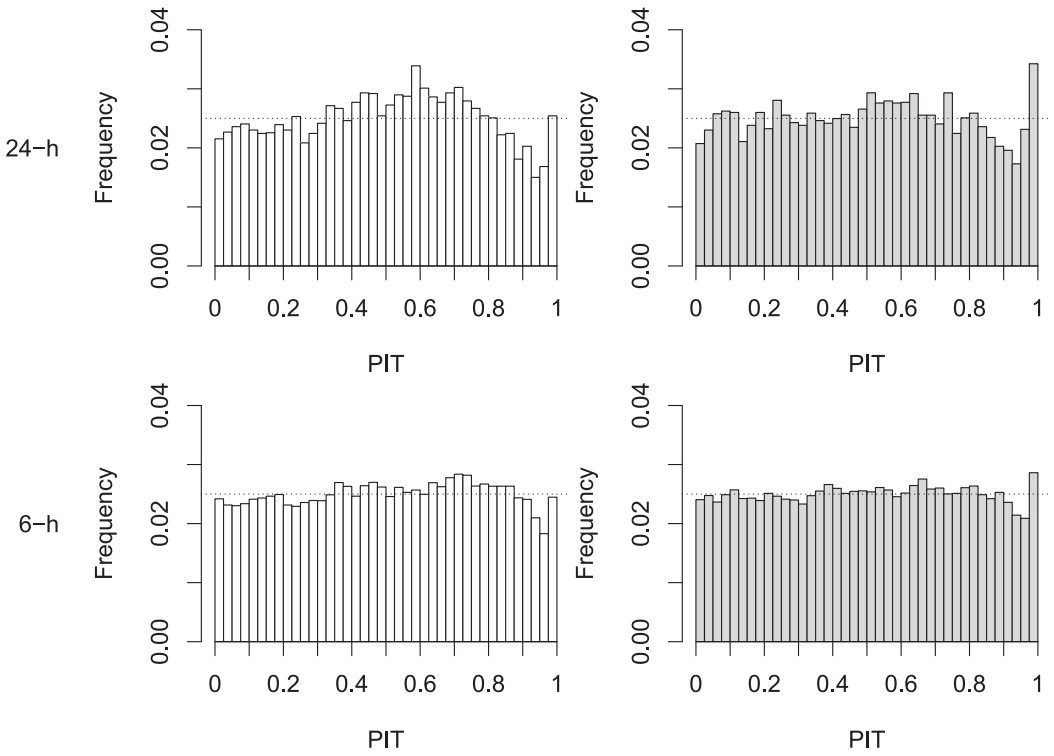

FIG. 7. The PIT analysis for log-link models with (left) censored Gaussian predictions and (right) censored logistic predictions, respectively. Results from tenfold cross validation over 10 stations and different lead times (+24 to +48 h), are evaluated separately for different accumulation periods (24; 6 h). The bin width is 0.025.

large. As a result, BS values are high if a corresponding event did not occur.

Brier scores of probability forecasts are lower (better) for high thresholds for both accumulation periods. Such events are rare as this threshold is based on the observed 97% quantile. The raw ensemble seems to already capture these events quite well, which is indicated by the similar BS as the statistical models (Table 2). Nevertheless, the statistical models improve BS values for 24-h sums and 6-h sums by about 3% and 6%, respectively (Table 3). A slightly stronger improvement on 6-h sums can be seen for the logistic models than for the Gaussian models, which highlights the importance of the heavy tailed distribution, further discussed in section 4c.

b. *Split approach and split levels*

After having shown an improvement against the raw ensemble particularly in PoP, but also in CRPS and the high BS threshold, we will in the following focus on the improvements from the statistical refinements and start with the split approach.

Figure 5 summarizes CRPSS values for censored Gaussian and logistic models, relative to our reference approach where the squared-scale parameter is optimized without additional information of members being zero (quad). The boxplots represent individual cases (lead times) for each station, which are between the +24- and +48-h forecast lead times; 24-h sums include 10 CRPSS values, and 6-h sums include 40 CRPSS values.

The forecast skill increases for all split models using $\nu = 1$, which is even more pronounced for the 6-h accumulation periods. Median values are highest for split models using the log link and identity link. This pattern is similar for censored logistic models, especially for 6-h sums. Additionally, Table 2 displays the smallest median CRPS values for the log-split models on 6-h sums.

Figure 6 compares different split levels ($\nu = 0.02$, 0.1, 0.5) and shows CRPSS in reference to a model with $\nu = 1$. Here only the log link is used but results look similar for other link functions as well. A split level of 0.02 is clearly worse but a split level of 0.5 performs almost equally compared to the reference. This result is the same for censored Gaussian and censored logistic split models using the log link.

However, an optimum split level, which can be found by testing all possible levels on training data, indicated only a small CRPSS improvement for 6-h sums against the split level of $\nu = 1$ (result not shown). Optimum values for ν range from 0.02 to 0.71 and 0.02 to 1 for 24- and 6-h sums, respectively.

TABLE 4. Median values for Brier score and its decomposition for thresholds of precipitation amounts larger than quantiles q0 (0 mm) and q97 of observed amounts for both accumulation periods. Columns show Brier score (BS), reliability (REL), resolution (RES), and sharpness (SHARP) for accumulation periods of 24 and 6 h. Rows show values for Gaussian and logistic models. Values are based on a tenfold cross validation including all available forecasts (stations and lead times) for the investigated accumulation periods of forecast day 2. Binning used in this decomposition is based on 10% intervals.

		BS		REL		RES		SHARP	
		24 h	6 h	24 h	6 h	24 h	6 h	24 h	6 h
				Quantile q0					
Gaussian models	quad	0.1071	0.0868	0.0112	0.0054	0.1265	0.0571	2.9896	1.9146
	quad_frac	0.1062	0.0858	0.0092	0.0047	0.1243	0.0567	2.9654	1.9158
	quad_split	0.1069	0.0868	0.0110	0.0054	0.1255	0.0572	2.9884	1.8750
	log_split	0.1072	0.0853	0.0095	0.0055	0.1261	0.0599	2.9112	1.8708
	id_split	0.1059	0.0861	0.0092	0.0051	0.1262	0.0576	2.9842	1.8919
Logistic models	quad	0.1022	0.0842	0.0058	0.0037	0.1258	0.0576	3.1275	2.0239
	quad_frac	0.1021	0.0837	0.0055	0.0033	0.1262	0.0586	3.1177	2.0231
	quad_split	0.1021	0.0842	0.0059	0.0040	0.1259	0.0576	3.1231	2.0112
	log_split	0.1025	0.0831	0.0053	0.0039	0.1267	0.0600	3.0413	1.9946
	id_split	0.1014	0.0844	0.0054	0.0036	0.1273	0.0586	3.1253	2.0146
				Quantile q97					
Gaussian models	quad	0.0188	0.0232	0.0042	0.0050	0.0133	0.0095	1.1222	1.1190
	quad_frac	0.0187	0.0227	0.0045	0.0045	0.0132	0.0098	1.1303	1.0857
	quad_split	0.0187	0.0231	0.0042	0.0048	0.0132	0.0095	1.1303	1.1077
	log_split	0.0183	0.0226	0.0038	0.0044	0.0124	0.0097	1.0955	1.0956
	id_split	0.0185	0.0230	0.0046	0.0047	0.0136	0.0099	1.1284	1.1018
Logistic models	quad	0.0189	0.0225	0.0032	0.0043	0.0125	0.0092	1.0220	1.0138
	quad_frac	0.0190	0.0224	0.0032	0.0039	0.0121	0.0093	1.0226	0.9654
	quad_split	0.0189	0.0225	0.0033	0.0043	0.0125	0.0096	1.0220	1.0053
	log_split	0.0188	0.0220	0.0036	0.0037	0.0121	0.0093	1.0357	0.9951
	id_split	0.0189	0.0223	0.0036	0.0043	0.0121	0.0093	1.0234	0.9991

c. *Heavy tails*

Calibration is one of the most important properties of probabilistic forecasts. We therefore compute the probability integral transform (PIT), which is similar to rank histograms (Hamill and Colucci 1998; Anderson 1996; Talagrand et al. 1997). It bins the forecasted cumulative probability density function and counts into which bin the observed value falls. If the model is well calibrated the bins should all have the same number of observations.

Figure 7 shows PIT histograms for different accumulation periods. For simplicity, only split log-link models are shown, since they performed best in terms of 6-h CRPSS values. The remaining models generate very similar histograms (results not shown). Both, Gaussian and logistic models are better calibrated if short accumulation periods are forecasted. Logistic models generally produce histograms that are more uniformly distributed.

This improvement by the logistic tail is also quantified in terms of CRPS and BS values, summarized in Table 2. Shown are median CRPS and BS values for the second forecast day (+24 to +48 h) including all available stations. Additionally, the BS values are decomposed based on Murphy and Winkler (1987) for two thresholds. The BS and its probabilistic attributes of reliability, resolution, and sharpness are summarized in Table 4 for two thresholds of precipitation amounts (PoP, high). The BS values are similar among the models and decrease for short observation intervals in general. This is related to the number of zeros, which increases for shorter accumulation periods. Censored logistic models are better than censored Gaussian ones for PoP. The decomposition also shows a smaller reliability, slightly larger resolution, and larger sharpness of censored logistic forecasts. This logistic tail seems to better represent the observed distribution, which is indicated by the smaller value of reliability. Furthermore, the increased resolution indicates that the logistic tail better discriminates between precipitation and no-precipitation events. Hence, the sharpness is also larger because of this better distinction.

This logistic benefit decreases slightly for high thresholds of the 0.97 quantile with even fewer events. Although reliability is still improved, resolution becomes worse than for censored Gaussian models. This indicates that similar probabilities for the high threshold are forecasted, which is also displayed by the smaller sharpness. In terms of 24-h sums this also leads to slightly worse BS values.

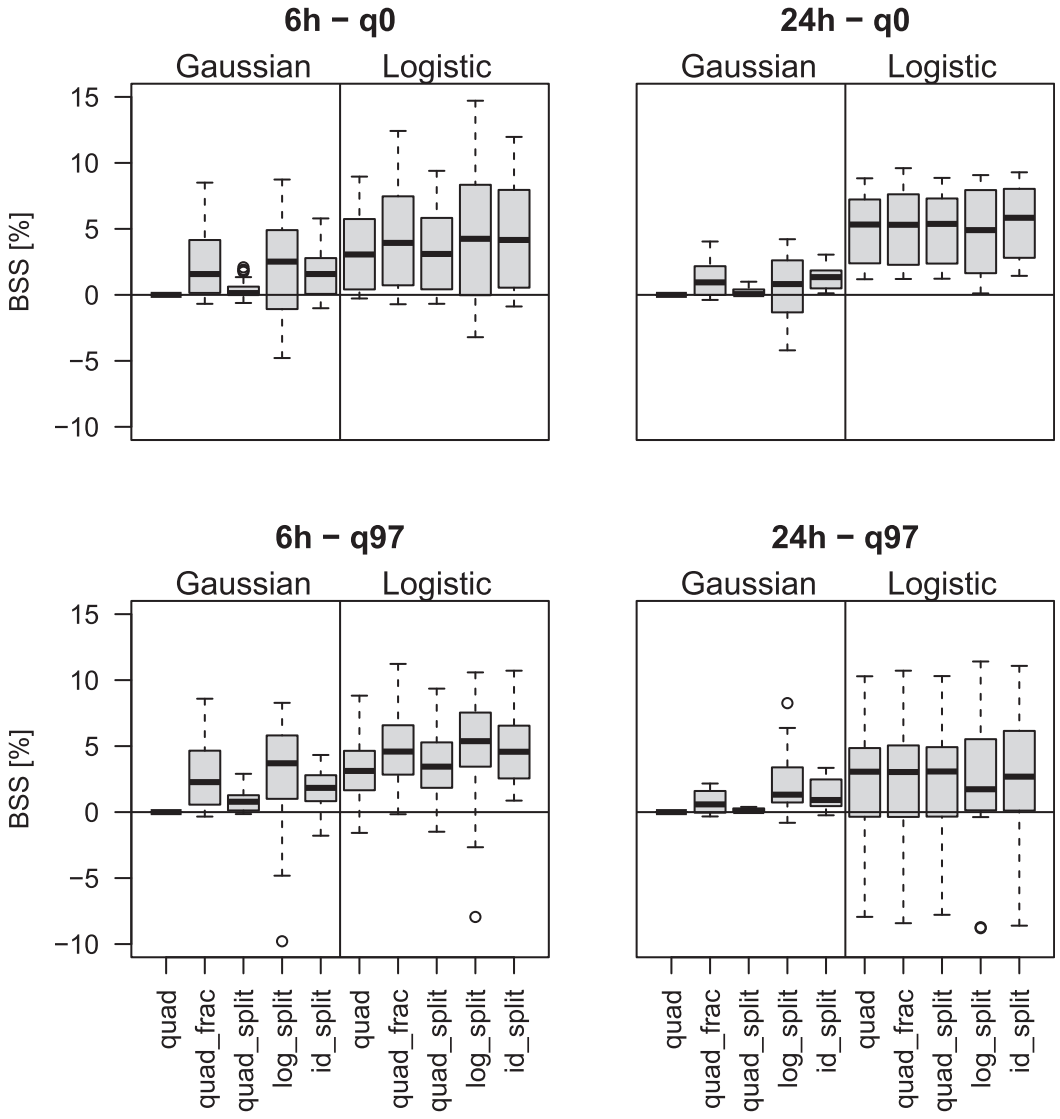

FIG. 8. BSS for censored Gaussian and logistic models (Table 1), shown for (left to right) 6- and 24-h sums. (from top to bottom) Thresholds for precipitation amounts larger 0 mm (q0) and the 97% quantile (q97) are used. Reference is the censored Gaussian quad model without splitting. Results are from tenfold cross validation, where each value represents individual lead times for 10 stations, and lead times are between +24 and +48 h. Boxplots are as in Fig. 5.

A verification of individual stations and lead times regarding BSS is illustrated in Fig. 8. A clear skill by the logistic tail is visible for PoP of 24-h sums. This also agrees with smaller reliability values of Table 4 and the visible calibration of Fig. 7. Similarly, the medians show an improvement for the high threshold (24 h–0.97 quantile) and 6-h sums for both thresholds against the baseline approach. Regarding the high threshold, individual cases show a negative skill for 24-h sums.

Thresholds even higher than the 97 percentile have not been investigated due to the insufficient number of events used for binning.

d. Link functions

So far we have seen an improved skill by using split models and the logistic tail. CRPS differences in the split models (Fig. 5) might be understood by looking at the regression fits for different link functions. Figure 9 gives an example fit for censored logistic models for cases where the standard deviation of the raw NWP ensemble model was larger than 0. Results for this example at Innsbruck for +36 h display a general pattern that can be found for the entire study area. While the linear fits for the latent mean value (left graphic) do not vary much,

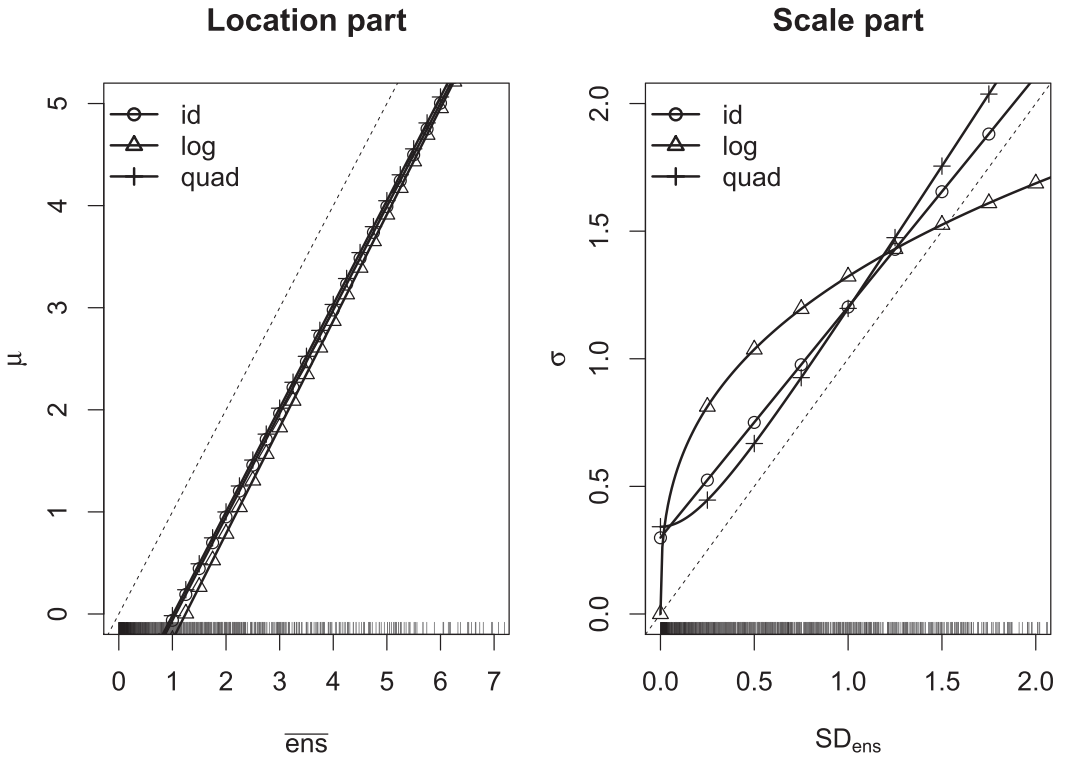

FIG. 9. Link functions for censored logistic models showing predictions of location (latent mean) and scale at Innsbruck, 6-h sum for lead time +36 h: identity link (circle), log link (triangle), quadratic link (cross); *x* axis denotes the ensemble mean $\overline{\text{ens}}$ for location models and the ensemble standard deviation SDens for scale models. The rug bars on the *x* axis illustrate the raw ensemble values used for fitting.

the fit for the scale parameter (right graphic) highlights larger differences. If the ensemble was already perfect, the fitted curves would follow the dashed black line. Since this is not the case, ensemble mean values are corrected to lower values (fits below the black line) and ensemble standard deviations are corrected to higher values (fits above the black line).

Differences in the predicted scale parameter can be seen especially for small values of the ensemble standard deviation (e.g., $\text{SD}_{\text{ens}} = 0.5$), where the log-link predictions are largest. Furthermore, the log-link model predicts smallest values of the latent mean value for small ensemble mean values. If the latent mean becomes more negative, the scale parameter has to be larger in order to still capture the observations.

This pattern reverses at a certain point, where the quad-link produces the largest scale parameters.

e. *Lead time performance*

Previously shown results are based on forecasts for day 2 (+24 to +48 h). To investigate lead time performance, the comparison setup is extended on station Innsbruck where additional interpolated NWP data are available up to +144 h.

Figure 10 displays CRPSS values for 6-and 24-h sums from +24 to +144 h. CRPSS values are shown for the censored logistic split model using the log link, which performed best in the previous analysis. To illustrate the combined performance of all three statistical refinements, the skill score reference is the censored Gaussian model using the quadratic link. The bootstrapped CRPSS values in Fig. 10 clearly show an improvement over all lead times for both accumulation periods. A stronger decay in forecast performance is visible after day 2 but the three refinements still improve the forecast performance up to +144 h.

These improvements result from a combination of the three proposed refinements. Also, calibration evaluated over all lead times is found to be similar to the 2-day lead time (Fig. 7). As lead time increases, the number of unanimous ensemble forecasts where *all* members have either precipitation or no precipitation decreases, indicating that the ensemble is generally less certain about precipitation occurrence (Fig. 11). As a result the effectiveness of the split approach also decreases, so that the remaining improvements should likely be ascribed to the log-link and the logistic distribution, where the heavier tail of

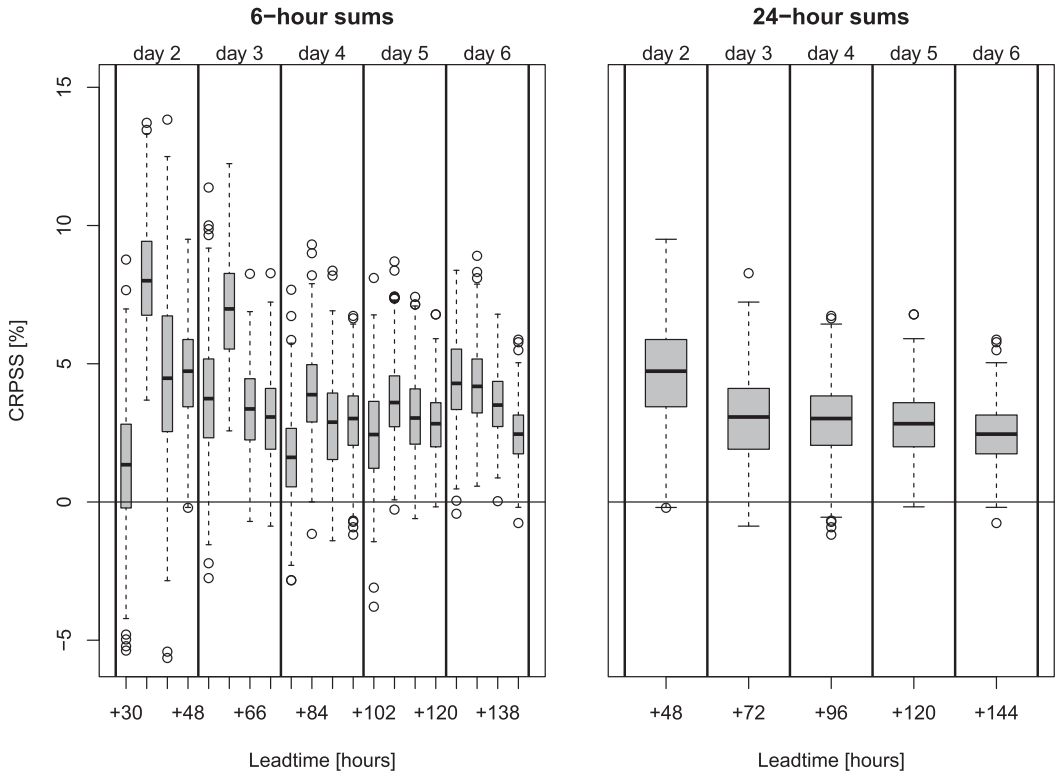

FIG. 10. CRPSS for the censored logistic model using the split approach and the log link (log_split) in reference to the censored Gaussian model using the quadratic link without split (quad). Results from tenfold cross validation are evaluated separately for lead times +30 to +144 h for (left) 6-h sums and (right) 24-h sums for Innsbruck. Each lead time contains 500 bootstrapped CRPSS values. Boxplots are as in Fig. 5.

the logistic distribution is more pronounced (results not shown).

5. Summary and conclusions

In this study we set out to investigate the effects of three refinements to a nonhomogeneous regression (NHR) model (e.g., Gneiting et al. 2005; Thorarinsdottir and Gneiting 2010; Messner et al. 2014a; Scheuerer 2014) for precipitation forecasts (24-h and 6-h sums), and the probability of precipitation exceeding two thresholds (0 mm and 97% quantile of observations). The initial precipitation forecasts are based on the global ECMWF ensemble with 32-km horizontal grid size. Namely, we propose a split approach to exploit unanimous zero precipitation ensemble predictions, use a heavy tailed distribution to better describe the tail behavior of precipitation data, and assess various link functions to ensure nonnegativity of the predictive variance. A case study on 10 sites in a small study area in the European Alps shows that especially the split approach can clearly improve the predictive performance for 6-h sums.

The split approach can exploit the fact that in our dataset unanimous zero precipitation ensemble predictions almost always perfectly predict dry events. By switching to a different model parameter set for these situations in the statistical models, the forecast performance can clearly be improved. The approach also allows us to relax ''unanimous'' to ''majority'' of ensemble members, and ''no precipitation'' to precipitation not exceeding other thresholds. Such modifications did not improve results reported here but might be beneficial for other datasets with different precipitation climatologies.

Furthermore, using the censored logistic distribution increases the forecast skill compared to censored Gaussian models. The pronounced tail of the logistic distribution is able to better capture rare events and improved CRPS, calibration in terms of PIT, and BS values. Regarding the probability of high precipitation amounts, the logistic models can improve reliability but showed a lack of resolution. The refinement of a better parametric representation of the precipitation distribution might also be accomplished with other distributions such as the (censored shifted) Gamma or the generalized

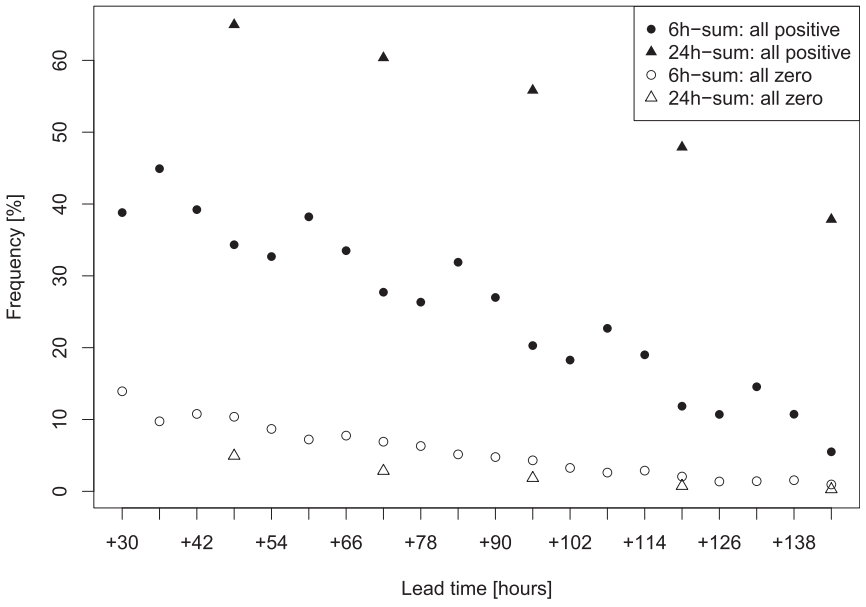

FIG. 11. Frequency of the 51-member ECMWF ensemble unanimously forecasting no precipitation (unfilled symbols) or precipitation (filled symbols) evaluated at Innsbruck for accumulation periods of 24 (triangles) and 6 h (circles).

extreme value distribution (Wilks 1990; Scheuerer 2014; Scheuerer and Hamill 2015).

Our third refinement is the investigation of different link functions for the dispersion submodel in the NHR approach. Depending on the used forecast distribution, distribution parameters may require positive values during numerical optimization. We find notable differences in forecast skill for different link functions, especially for short accumulation periods of 6-h sums. The best performance has been achieved by using the log link, which optimizes the logarithmic standard deviation (Messner et al. 2014a). This approach outperformed simulations where the squared scale (quad) or scale parameter (id) is estimated. Although all link functions could correct for the underdispersion of the raw ensemble, id and quad models often still have too little uncertainty for smaller ensemble standard deviations, which occur most frequently. Additionally, the combination of log-link and the split approach allows us to use the logarithmic standard deviation as regressor. Otherwise, the logarithmic standard deviation could not be used directly due to infinity occurring for unanimous ensembles (zero standard deviation).

Although general verification has focused on day-2 forecasts, the proposed refinements perform similarly for lead times +30 to +144 h at one example station. Combining all three refinements yields an improved forecast performance compared to the baseline approach. Nevertheless, for the proposed split approach to be most effective, there have to be unanimous zero precipitation forecasts in the ensemble which usually occur more frequently at shorter lead times, shorter accumulation periods, and in regions with less precipitation events in general. Similarly, the used link function seems to be more influential for cases where generally smaller precipitation amounts occur. Hence, differences in link functions are found to be largest for short accumulation periods where amounts are usually smaller than for long periods. Conversely, the heavy tail of the censored logistic distribution is found to be more important for the longer accumulation periods (24-h sums), where precipitation amounts are generally higher.

The proposed refinements are not restricted to the presented NHR method but can also be combined with other extensions. Such extensions can cover the consideration of neighboring grid points (Scheuerer 2014), the use of additional information from high-resolution models (Hemri et al. 2016), or copula coupling approaches to ensure spatial correlation between investigated stations (Feldmann et al. 2015). Clearly, the effectiveness of all refinements strongly depends on the dataset and the area of interest. However, as the refinements do not require the acquisition of additional input data from NWP models, they are straightforward to apply and thus practitioners can easily check whether they lead to improvements for their data and study area. The key improvement appears to be the inclusion of

unanimous zero precipitation forecasts, especially at short(er) aggregation periods and lead times where the ensemble is typically more certain. Consequently, the refinements are expected to be valuable also for high-resolution ensemble systems.

To summarize the overall forecast performance for our study area, all statistical models could clearly improve the raw ensemble forecasts. Our results imply that an untransformed censored logistic assumption is adequate particularly for short accumulation periods (6-h sums). The split approach improves the forecasts by using the information of zeros predicted by the raw ensemble. Results also showed differences in link functions where the logarithmic link performed best. Together, our three statistical refinements provide the largest benefits for short accumulation periods (6 h) and short lead times.

Acknowledgments. This work is an ongoing Ph.D. project and was funded by Autonome Provinz Bozen-Abteilung Bildungsförderung, Universität und Forschung (ORBZ110725). We thank the Austrian weather Service (ZAMG) for their supply of ECMWF EPS data, and the weather service of South Tyrol for their observational data. Many thanks also to Reto Stauffer, who did a lot of the data handling.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecast from ensemble model integration. *J. Climate*, **9**, 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, doi:10.1038/nature14956.

Ben Bouallègue, Z., and S. E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: From verification results to probabilistic products. *Meteor. Appl.*, **21**, 922–929, doi:10.1002/met.1435.

Bentzien, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002, doi:10.1175/WAF-D-11-00101.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Carisse, O., R. Bacon, A. Lefebvre, and K. Lessard, 2009: A degree-day model to initiate fungicide spray programs for management of grape powdery mildew [*Erysiphe necator*]. *Can. J. Plant Pathol.*, **31**, 186–194, doi:10.1080/07060660909507592.

Cohen, A. C., 1959: Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, **1**, 217–237, doi:10.1080/00401706.1959.10489859.

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 2015: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, **143**, 955–971, doi:10.1175/MWR-D-14-00210.1.

Gneiting, T., and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151, doi:10.1146/annurev-statistics-062713-085831.

——, A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724, doi:10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2.

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, doi:10.1256/qj.06.25.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, doi:10.1175/MWR-D-15-0004.1.

Hemri, S., T. Haiden, and F. Pappenberger, 2016: Discrete post-processing of total cloud cover ensemble forecasts. *Mon. Wea. Rev.*, **144**, 2565–2577, doi:10.1175/MWR-D-15-0426.1.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Klein, N., T. Kneib, S. Lang, and A. Sohn, 2015: Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.*, **9**, 1024–1052, doi:10.1214/15-AOAS823.

Krzysztofowicz, R., and A. A. Sigrest, 1999: Calibration of probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **14**, 427–442, doi:10.1175/1520-0434(1999)014<0427:COPQPF>2.0.CO;2.

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, doi:10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Löpmeier, F., K. Wittich, C. Frühauf, and S. Schittenhelm, 2012: Entwicklungen und Stand der Aktivitäten in der Agrarmeteorologie (Development and current level of activities in agrometeorology). *Promet*, **38**, 2–10.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, doi:10.1175/MWR-D-13-00355.1.

——, ——, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, doi:10.1175/MWR-D-13-00271.1.

——, ——, and ——, 2016: Heteroscedastic censored and truncated regression with crch. *R J.*, **8**, 173–181, https://journal.r-project.org/archive/2016-1/messner-mayr-zeileis.pdf.

Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **17**, 173–191, doi:10.1175/1520-0434(2002)017<0173:TIOHRA>2.0.CO;2.

Murphy, A., and R. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338, doi:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

R Core Team, 2016: The R project for statistical computing. R Foundation for Statistical Computing, accessed 5 October 2016, http://www.r-project.org/.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, doi:10.3402/tellusa.v55i1.12082.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, doi:10.1002/qj.2183.

——, and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, doi:10.1175/MWR-D-15-0061.1.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/MWR3441.1.

Smith, R. B., Q. Jiang, M. G. Fearon, P. Tabary, M. Dorninger, J. D. Doyle, and R. Benoit, 2003: Orographic precipitation and air mass transformation: An Alpine example. *Quart. J. Roy. Meteor. Soc.*, **129**, 433–454, doi:10.1256/qj.01.212.

Stauffer, R., G. J. Mayr, J. W. Messner, N. Umlauf, and A. Zeileis, 2017a: Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *Int. J. Climatol.*, **37**, 3264–3275, doi:10.1002/joc.4913.

——, N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017b: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Wea. Rev.*, **145**, 955–969, doi:10.1175/MWR-D-16-0260.1.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability*, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268, doi:10.1017/S1350482705001763.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, doi:10.1111/j.1467-985X.2009.00616.x.

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate*, **3**, 1495–1501, doi:10.1175/1520-0442(1990)003<1495:MLEFTG>2.0.CO;2.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, doi:10.1002/met.134.

——, 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, doi:10.1175/MWR3402.1.

Zhu, Y., and Y. Luo, 2015: Precipitation calibration based on the frequency-matching method. *Wea. Forecasting*, **30**, 1109–1124, doi:10.1175/WAF-D-13-00049.1.